Approaches to sentiment analysis of the social network text data

Vadim Moshkin
Ulyanovsk State Technical University
Ulyanovsk, Russia
v.moshkin@ulstu.ru

Nadezhda Yarushkina
Ulyanovsk State Technical University
Ulyanovsk, Russia
jng@ulstu.ru

Ilya Andreev
Ulyanovsk State Technical University
Ulyanovsk, Russia
ia.andreev@ulstu.ru

***Abstract*—The article provides an overview of the most modern approaches to sentiment analysis of text data. The features of using machine learning approaches and dictionary-based methods are also described. In addition, the description of sentiment dictionaries and the most popular software for sentiment analysis of data are given. An original approach was also proposed for sentiment analysis of text data using the integration of machine learning methods with the Wodr2vec data vectorization algorithm. Also presented is the architecture of the developed system for Opinion Mining data of social networks. At the end of the article, experiments are presented to evaluate text reviews using the data from the IMDB portal as an example, confirming the proposed approach.**

Keywords—sentiment analysis, word2vec, Opinion Mining, machine learning

I. Introduction

Currently, the main source of information from where you can get knowledge about certain interests of the client, prepare for him and proactively offer a new product or service, are the Internet and social networks [1]. This problem is solved by the Opinion mining. Opinion mining for data from social networks contains two tasks:

- morphological analysis to identify entities that will be evaluated;
- analysis of the sentiment of expressions related to this entity.

By sentiment analyzing of the users' text messages the researcher can draw conclusions about:

- emotional evaluation of users of various events and objects;
- individual user preferences;
- some features of the users' nature [2].

Sentiment analysis is a section of text mining, a system for automatically extracting subjective opinions from text. Sentiment analysis determines the content of the text as much as its tonality.

Automatic analysis of the tonality of the text is based on the technologies of linguistic interpretation of emotions, machine learning, extracting emotional meaning from information, etc.

The technology of sentiment analysis has become especially relevant with the development of Web 2.0, as a tool for monitoring the views of millions of Web users.

However, text data in social networks have the following features:

- use whole and incomplete sentences.
- the presence of speech and spelling errors.
- the use of smiles, emoji to give the message a certain emotional coloring.

In this article we will consider the use of various existing algorithms for assessing the sentiment of social network texts within the framework of the developed software system for Opinion Mining. The article proposes an original approach for analyzing the emotional coloring of text data using the integration of machine learning methods with the Wodr2vec algorithm.

II. The existing methods and software for sentiment analysis of text data

There are two main groups of methods for the automatic sentiment analysis of text data:

A. Statistical methods

The basis of these methods is the use of machine classifier. This classifier is learned on pre-marked texts in the first stages. Then the classifier builds a model for analyzing new documents using the knowledge gained. The algorithm consists of:

- a collection of documents is collected for machine classifier learning;
- each document is decomposed into a feature vector;
- the correct sentiment type is indicated for each document;
- the selection of the classification algorithm and the method for learning the classifier;
- the resulting model is used to determine the documents sentiment of the new collection.

The disadvantage of such methods is the need for a large amount of data for learning.

The statistical approach widely uses the support vector method (SVM) [3], Bayesian models [4], various types of regression [5], methods Word2Vec, Doc2Vec [6], CRF [7], convolutional and recurrent neural networks [8][9].

Word2Vec. The Word2Vec method is based on the vector representation of words and the determination of the semantic proximity of lexical units based on their distribution in collections of texts on specific topics.

A big set of texts are input to Word2Vec. Specialized vocabulary is created and at the same time is learned on this set of texts. At the second stage, the dictionary turns into a set of vector representations of words. This representation is based on the contextual proximity of a given word: if the

words are found in the text side by side often enough, then there is a semantic connection between them, and therefore, in the vector representation, these words will have close coordinates.

For this algorithm, two training methods were developed - CboW and Skip-gram. Schemes of these algorithms are presented in Figure 1. The first algorithm is based on the prediction of the next word in the sequence given during training. The second learning method works differently - it predicts the surrounding words. The result of this method is the ability to calculate the "semantic distance" for each pair of words.

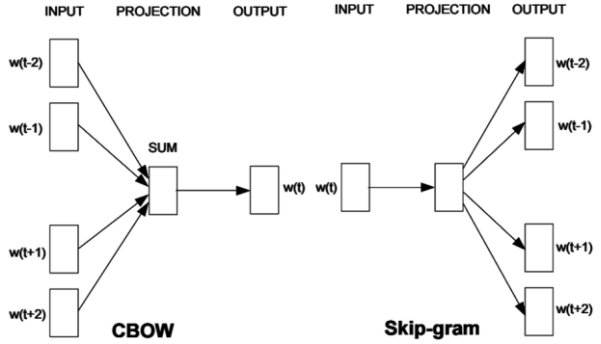

Fig. 1. CboW and Skip-gram training methods.

Doc2Vec. The Doc2Vec method consists of two methods: distributed memory (DM) and distributed word bag (DBOW). The DM method predicts a word from known prior words and a paragraph vector. The paragraph vector does not move and takes into account the word order Despite the fact that the context moves through the text. DBOW predicts random word groups in a paragraph based only on the paragraph vector.

A serious disadvantage of this method is the complexity of the analysis of the training sample, which is why it is extremely difficult to continuously update the model when new training data is received.

B. Methods based on dictionaries

The method using dictionaries is based on the search for emotive vocabulary (lexical tonality) in the text according to pre-compiled tonal dictionaries and rules using linguistic analysis.

These methods can use rule lists that are substituted into regular expressions and special rules for connecting tonal vocabulary within sentences [10].

Glossary terms must have a weight corresponding to the subject area of the document in order to classify the document with high accuracy. Emotion is taken into account in the algorithm when finding the marker. The result of the algorithm is the average emotional color of the text [11-12]. The following algorithm is usually used:

- assign the sentiment score from the dictionary to each word in the text;
- calculate the overall sentiment score of the entire text by adding the sentiment score of individual words [13].

The disadvantage of this method is a significant amount of labor because the method requires the creation of many rules.

A mixed method is also sometimes used [14-16].

C. Dictionaries and thesauri

There are a number of thesauri labeled with regard to the emotional component. These dictionaries are necessary for computer programs when analyzing the tonality of the text.

WordNet-Affect is a semantic thesaurus in which concepts related to emotions are represented using words that have an emotional component. WordNet-Affect also uses additional emotional labels to separate synsets according to their emotional valency. To do this, four additional emotional labels are defined:

- positive;
- negative;
- ambiguous;
- neutral.

SentiWordNet [17] is a lexical semantic thesaurus. The first version of SentiWordNet was developed in 2006. This thesaurus appeared as a result of automatic annotation of each set of synonyms in accordance with its degree of positivity, negativity and objectivity.

SenticNet is another semantic thesaurus for working with sets of emotional concepts. SenticNet is used to design intelligent applications designed to analyze the emotional component of text. The main purpose of SenticNet is to simplify the process of machine recognition of conceptual and emotional information transmitted using natural language. If we compare other lexical thesauruses, such as SentiWordNet and WordNet-Affect with SenticNet, then their main difference is that SentiWordNet and WordNet-Affect provide the linking of words and emotional concepts at the syntactic level, not allowing to reveal the semantic component.

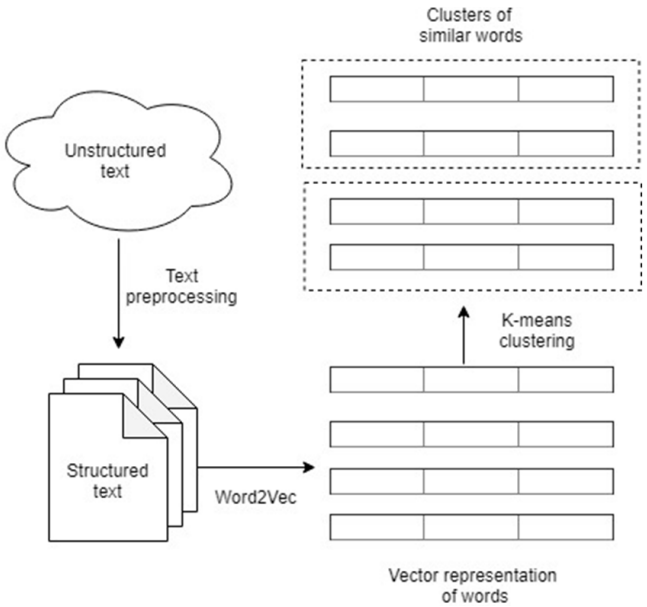

Fig. 2 Developed clustering algorithm.

D. *Existing sentiment analysis software.*

Currently there is a certain set of libraries and software for sentiment analysis of text data.

Chorus is a service for determining the emotional coloring of email. This service was a startup and was developed by a company from Australia. Chorus is intended for customer support services:

- recommends the following message for processing;
- indicates a message that needs an urgent response;
- indicates where you can save the client after the response.

The disadvantage is the ability to analyze only emails. Currently no longer supported.

Sentiment Analysis with Python NLTK Text Classification [18] is a demo showing the capabilities of NLTK. He divides the emotional coloring into positive, negative and neutral. An API with restrictions and the ability to buy premium access is also offered. The demo sample is a form for manual verification with character size restrictions.

Sentirength [19] is a library for analyzing emotional coloring. The algorithm is based on the search for the maximum tonality value in the text for each scale (ie, the search for the word with the maximum negative rating and the word with the maximum positive rating) [20]. As a result, a double score (positive and negative) is given from 1 to 5. There are also options for triple and single assessment of results. This library is paid. You can check the library on the project website.

Tone Analyzer [21] is a service from IBM based on IBM Watson. This service uses linguistic analysis to detect emotional and linguistic connotations in the written text. Options for using the analyzer are social listening, improving the quality of customer service and integration with chat bots. This service is paid and supports only English.

III. Sentiment Analysis using machine learning and Word2Vec.

The Random forest method of text sentiment analysis is a clustering method based on machine learning.

Schematically, the developed algorithm is presented in the Fig.2.

1) Text data pre-processing is carried out at the first stage. The html code, any non-alphabetic characters, and also stop words are removed from the text. Stop words are phrases and words that do not carry a semantic load and make it difficult to index a page by search engines. Further, all remaining words are reduced to lowercase.

2) At the second stage, the text from these files (test and training) presented in the form of a list of significant words is processed using the Word2Vec tool.

Word2vec is an open source tool for calculating word spacing provided by Google [22]. Word2Vec creates a special model that includes a dictionary of words with their vector representation.

By the similarity of the values of the vectors, synonyms and similar words can be determined. 300-dimensional vectors were used to most accurately identify words. The resulting model is saved as a file.

3) The next stage is the clustering of vector words according to the K-Means method for splitting by synonyms and similar words. The number of clusters should be such that on average there are 5 words per cluster, for the most accurate result.

It is required to prepare data for machine learning after breaking all the significant words into clusters. A two-dimensional array is created for each file as follows:

- the number of lines is equal to the number of text messages in the file;
- the number of columns is equal to the number of clusters.

These data will be important in determining the emotional coloring.

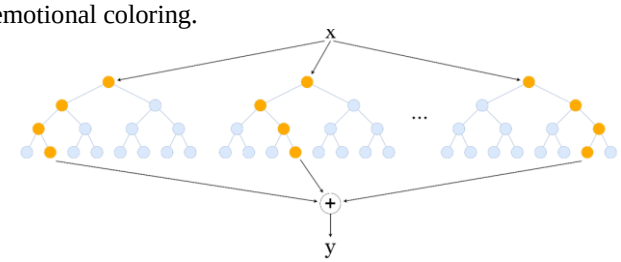

Fig. 3 Random Forest Sentiment Analysis of the text.

The Random Forest model (Fig. 3) was used for machine learning. The random forest method is currently one of the most popular and effective methods for solving machine learning problems, such as classification and regression. He trains not one decision tree with his weights, but many decision trees [23].

Predicting data and calculating the accuracy of the algorithm is performed using a trained model.

IV. Software Architecture for Opinion Mining Social Media

A module for assessing the tonality of texts in the information system for Opinion Mining (Fig.4) was developed to evaluate the effectiveness of the proposed algorithms [24-26].

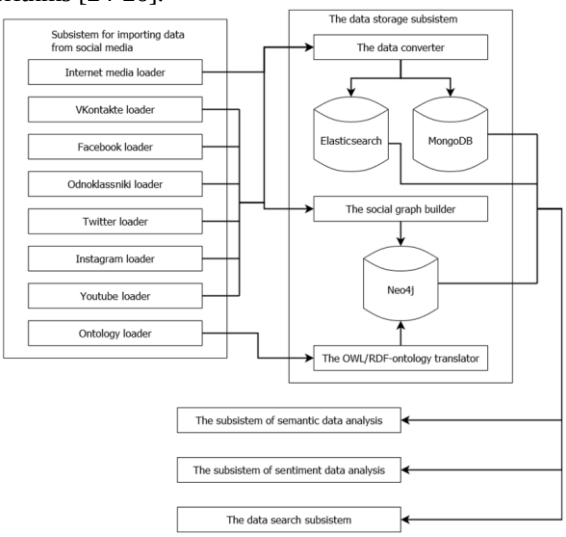

Fig. 4. The architectural scheme of the software system for Opinion Mining social media.

This information system solves the following tasks:

- extracts data from various social networks (Facebook, Ok, VKontakte, Instagram, Twitter, etc.)
- conducts preprocessing of the extracted data;
- makes matching (comparison) of user profile data from different social networks;
- translates the extracted data into an internal format for storing knowledge;
- conducts semantic analysis of data using subject ontologies to simplify the search;
- conducts sentimental analysis of the extracted data using the developed algorithm.

The developed software system for Opinion Mining has a service architecture and supports the REST architectural style. The ElasticSearch library is used to extract and preprocess data. MongoDB is used to store a large set of data. The Sypher query language is used to search the graph database Neo4j [27-28].

V. Experiment results.

Experiments were conducted to determine the accuracy of estimating the emotional coloring of text data using the random forest method.

Test data is a data set from the IMDB site that contains 100,000 detailed film reviews (positive and negative). 1,500 reviews were taken separately to verify accuracy. The maximum accuracy is 79% because some reviews do not contain emotional coloring, but are only a retelling of the plots of films, which lowered the accuracy of the program.

When using different parameters, the running time of the algorithm ranged from 40 to 55 minutes. In the experiments, the optimal values of the algorithm's work were revealed, such as the dimension of the vectors, the number of clusters and the minimum amount of use of the word in the reviews to make it important.

The results of the experiments are presented in Table 1 and Fig.5.

The best result was shown when using 300-dimensional vectors, the minimum number of repetitions of words equal to 60 and the number of vectors calculated so that each cluster had an average of 5 words, i.e. 3956 clusters.

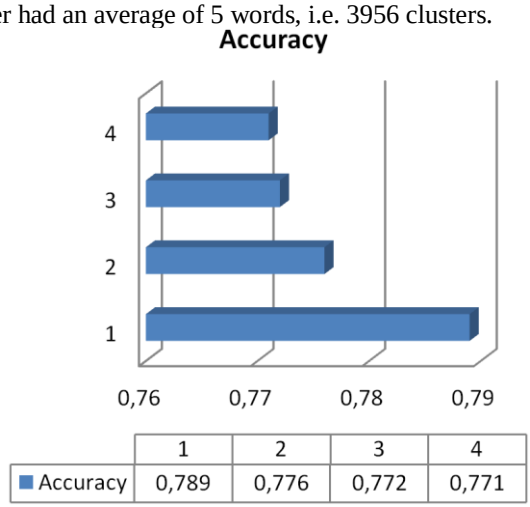

Fig. 5. Sentiment analysis using machine learning and Word2Vec.

Table I. Results of experiments

Accuracy	Number of vector spaces	Number of clusters	Min. number of words
0.789	300	3956	60
0.776	200	2301	100
0.772	400	2301	100
0.771	300	1573	60

Conclusion

Thus, in this paper, an approach to the analysis of the text data of social networks was proposed. This approach is based on the integration of the word2vec vectorization algorithm and the k-means clustering algorithm using a random forest algorithm for training a neural network. This approach was implemented in the Opinion Mining analysis system.

Experiments were conducted to evaluate the effectiveness of this algorithm when analyzing user feedback from the IMDB portal. The experiments showed that the Best result was shown using 300-dimensional vectors, the minimum number of repetitions of words was 60, and the number of vectors was calculated so that each cluster had an average of 5 words, i.e. 3956 clusters.

In future works, we plan to hybridize this approach using well-known sentimental ontologies and dictionaries to take into account the peculiarities of word usage and language..

Acknowledgment

This work was supported by the Russian Federal Property Fund. Projects No. 18-47-730035 and 18-47-732007.

References

[1] O. Shipilov and A. Belyaev, “Analysis of the emotional color of messages in the social network twitter,” Science Questions, vol. 3, pp. 91-98, 2016.

[2] D. Vlasov, “Description of the information image of a social network user, taking into account its psychological characteristics,” International Journal of Open Information Technologies, vol. 6. no. 4, 2018.

[3] M.S. Sabuj, Z. Afrin and K.M.A. Hasan, “Opinion Mining Using Vector Machine for Web Based Diverse Data,” Pattern Recognition and Machine Intelligence. Lecture Notes in Computer Science, vol. 10597, pp. 673-678, 2017.

[4] L.P. Dinu and I. Iuga, “The Best Feature of the Set,” Computational Linguistics and Intelligent Text Processing. Lecture Notes in Computer Science, vol 7181, pp. 556-567, 2012.

[5] I. Chetviorkin and N. Loukachevitch, “Sentiment Analysis Track at ROMIP-2012,” Computational linguistics and intellectual technologies: Sat scientific articles, vol. 2, pp. 40-50, 2013.

[6] Q. Chen and M. Sokolova, “Word2Vec and Doc2Vec in Unsupervised Sentiment Analysis of Clinical Discharge Summaries,” CoRR abs 1805.00352, 2018.

[7] A. Antonovam and A. Soloviev, “Using the conditional random fields method for processing texts in Russian,” Computational linguistics and intellectual technologies: Sat scientific articles, vol. 12, no. 19, pp. 27-44, 2013.

[8] A. Maas, R. Daly, P. Pham, D. Huang, A. Ng and C. Potts, “Learning word vectors for sentiment analysis,” The International Language Technologies. International Association for Computational Linguistics, vol. 1, pp. 142-150, 2011.

[9] Yu.V. Vizilter, V.S. Gorbatsevich and S.Y. Zheltov, “Structure-functional analysis and synthesis of deep convolutional neural networks,” Computer Optics, vol. 43, no. 5, pp. 886-900, 2019. DOI: 10.18287/2412-6179-2019-43-5-886-900.

[10] H. Saif, “Contextual semantics for sentiment analysis of Twitter,” Information Processing & Management, vol. 52, no. 1, pp. 5-19, 2016.

[11] A. Pak and P. Paroubek, “Twitter as a Corpus for Sentiment Analysis and Opinion Mining,” LREC, 2010.

[12] A. Tarasova, “Synergy of interrogative and exclamation marks in network texts (on the material of Tatar, Russian and English languages),” Bulletin of Vyatka State University, vol. 4, 2015.

[13] S. Ionova, “Emotiveness of a Text as a Linguistic Problem,” Diss .... Cand. filol. Sciences, 1998.

[14] B. Pang, L. Lee and S. Vaithyanathan, “Thumbs up?” Sentiment Classification using Machine Learning Techniques, pp. 79-86, 2002.

[15] P. Turney, “Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews,” Proceedings of the Association for Computational Linguistics, pp. 417-424, 2002.

[16] I.A. Rycarev, D.V. Kirsh and A.V. Kupriyanov, “Clustering of media content from social networks using BigData technology,” Computer Optics, vol. 42, no. 5, pp. 921-927, 2018. DOI: 10.18287/2412-6179-2018-42-5-921-927.

[17] V. Moshkin, N. Yarushkina and I. Andreev, "The Sentiment Analysis of Unstructured Social Network Data Using the Extended Ontology SentiWordNet," 12th International Conference on Developments in eSystems Engineering (DeSE), Kazan, Russia, pp. 576-580, 2019.

[18] Natural Language Processing APIs and Python NLTK Demos [Online]. URL: https://text-processing.com/demo/sentiment/.

[19] A. Esuli and F. Sebastiani, “SENTIWORDNET: A Guide for Respecting the Opthion Mining,” pp.417-422, 2006.

[20] SentiStrength [Online]. URL: http://sentistrength.wlv.ac.uk/.

[21] I. Menshikov and A. Kudryavtsev, “Review of systems for analysis of tonality of a text in Russian,” Young scientist, no. 12, pp. 140-143, 2012.

[22] Watson Tone Analyzer [Online]. URL: https://www.ibm.com/cloud/watson-tone-analyzer.

[23] Introduction to Word Embedding and Word2Vec [Online]. URL: https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa.

[24] J. Žižka, F. Dařena and A. Svoboda, “Random Forest,” 2019. DOI: 10.1201/9780429469275-8.

[25] N. Yarushkina, A. Filippov, M. Grigoricheva and V. Moshkin, “The Method for Improving the Quality of Information Retrieval Based on Linguistic Analysis of Search Query,” Artificial Intelligence and Soft Computing. Lecture Notes in Computer Science, vol. 11509, pp. 474-485, 2019.

[26] A. Pazelskaya and A. Soloviev, “Method for determining emotions in texts in Russian,” Computational linguistics and intellectual technologies: Sat scientific articles, vol. 11, no. 18, pp. 510-523, 2011.

[27] A. Filippov, V. Moshkin and N. Yarushkina, “Development of the Social Media Analysis,” Recent Research in Control Engineering and Decision Making. Studies in Systems, Decision and Control, vol. 199, pp. 421-432, 2019.

[28] N. Yarushkina, A. Filippov, V. Moshkin, G. Guskov and A. Romanov, “Intelligent Instrumentation for Opinion Mining in Social Media,” Proceedings of the II International Scientific and Practical Conference Fuzzy Technologies in the Industry, Ulyanovsk, Russia, pp. 50-55, 2018.